Int. J. Appl. Math. Comput. Sci., 2022, Vol. 32, No. 1, 81–94
DOI: 10.34768/amcs-2022-0007

A DATA ASSOCIATION MODEL FOR ANALYSIS OF CROWD STRUCTURE

M. SAMI ZITOUNI [a], ANDRZEJ ŚLUZEK [b,*]

[a] Department of Biomedical Engineering
Khalifa University of Science and Technology
PO Box 127788, Abu Dhabi, United Arab Emirates
e-mail: mohammad.zitouni@ku.ac.ae

[b] Institute of Information Technology
Warsaw University of Life Sciences (SGGW)
ul. Nowoursynowska 166, 02-787 Warsaw, Poland
e-mail: andrzej_sluzek@sggw.edu.pl

The paper discusses a non-deterministic model for data association tasks in visual surveillance of crowds. Using detection and tracking of crowd components (i.e., individuals and groups) as baseline tools, we propose a simple algebraic framework for maintaining data association (continuity of labels assigned to crowd components) between subsequent video-frames in spite of possible disruptions and inaccuracies in tracking/detection algorithms. Formally, two alternative schemes (which, in practice, can be jointly used) are introduced, depending on whether individuals or groups can be prospectively better tracked in the current scenario. In the first scheme, only individuals are tracked, and the continuity of group labels is inferred without explicitly tracking the groups. In the second scheme, only group tracking is performed, and associations between individuals are inferred from group tracking. The associations are built upon non-deterministic estimates of memberships (individuals in groups) and estimates obtained directly from the baseline detection and tracking algorithms. The framework can incorporate any detectors and trackers (both classical or DL-based) as long as they can provide some geometric outlines (e.g., bounding boxes) of the crowd components. The formal analysis is supported by experiments in exemplary scenarios, where the framework provides meaningful performance improvements in various crowd analysis tasks.

Keywords: data association, visual surveillance, crowd analysis, algebraic model.

1. Introduction

Crowd behavior understanding is one of important yet very challenging applications of intelligent surveillance systems. It can be critical in detecting anti-social behaviors, providing early warnings of security breaches, identifying patterns of group behaviors, etc. Sadly, the recent outbreak of COVID-19 pandemic adds novel dimensions to these applications. All these tasks have to be performed in dynamically evolving scenes where groups or individuals move, interact, merge, split, disperse, etc.

Detection and tracking of crowd components (individuals and their groups) are fundamental tools of vision-based crowd monitoring, and a lot of work has been done in these areas. Nevertheless, the required results are usually more complicated than direct outcomes of detectors or trackers, and include issues such as patterns of group splitting and merging, changes in group sizes and membership, duration of group existence, etc. Additionally, the actual performances of *state-of-the-art* detectors and trackers are usually imperfect in complex crowd scenes. Apart from natural degrading effects (e.g., weather conditions) we can also expect deterioration caused by crowd density, low resolution of individual silhouettes, multiple similar targets, crowd dynamics (e.g., vaguely defined boundaries of groups), etc.

Data association (i.e., consistent labeling of correspondingly the same crowd components over sequences of video-frames) plays a vital role in the above tasks, and it is of secondary importance if this could be achieved explicitly by tracking, (re-)detection, combination of both or by any alternative means.

*Corresponding author

Since, nevertheless, detectors and trackers remain the fundamental tools in vision-based crowd analysis, we argue that a model is needed to integrate them into a uniform framework optimizing the reliability of data association. The model should be applicable to any typical detectors and trackers of individuals and groups. To the best of our knowledge, no such model seems to exist even though complementary use of detection and tracking is reported in numerous works on machine vision (e.g., Park and Brilakis, 2016; Bochinski *et al*., 2018; Kasprzak *et al*., 2012).

Thus, as the main contribution of the paper, we propose a formal framework for integrating crowd detection and tracking results into a uniform data association model. The model is non-deterministic, i.e., it takes into account limited performances of detecting/tracking results which are often expressed probabilistically. The model consists of two alternative schemes exemplifying the opposite scenarios (in practice, both schemes can be applied simultaneously or supplementally) as follows:

1. In the first scheme, we assume that only individuals can be effectively tracked, i.e., groups are too vaguely defined or too sparse for tracking, and can be only detected (not necessarily reliably). Thus, group data associations (including splitting, merging, disappearance, etc.) should be derived from tracking individuals.

2. In the second scheme, tracking groups is considered practical, while individuals cannot be tracked (even though some of them can be detected, not always reliably) because of crowd density, poor resolution of images, etc. Then, data associations for individuals are derived from group tracking results.

The proposed framework can improve performances of crowd monitoring systems, particularly in scenarios where detection or tracking can be temporarily corrupted or disrupted (e.g., due to visual conditions). By using the proposed data association schemes, the disruptions and gaps in the crowd structure description can be easier rectified, and crowds can be monitored more smoothly.

To the best of our knowledge, the framework is the first attempt to convert results of detection and tracking (for groups and individuals) into a non-deterministic data association method for crowd components over sequences of video-frames. The proposed model can be combined with almost any detection/tracking algorithms (both traditional and DL-based). In the presented implementation, we employ just exemplary detectors and trackers.

In Section 2, we briefly overview selected background works related to the proposed framework, mainly works on tracking (and detecting) individuals and groups. It is shown that some of these methods actually use data association as supplementary tools, which in our opinion further justifies significance of the proposed model. Section 3 contains the formal mathematical specification of the framework. Some details of the low-level tools (detectors and trackers) used in the exemplary implementations are also provided there. Finally, Section 4 presents a number of experimental studies performed on publicly available benchmark datasets. The experiments focus on scenarios, where the proposed model can improve performances of various crowd analysis tasks. Section 5 concludes the paper.

2. Related background works

Vision-based crowd analysis has been developing at least fifteen years, and most of the proposed methods focus on detection and tracking individuals, either explicitly in the context of crowd motion analysis (Jacques *et al*., 2007; Garcia-Martin *et al*., 2017; Wang *et al.*, 2020a) or as a multi-target tracking task (see a recent survey by Ciaparrone *et al*. (2020)). Nevertheless, the alternative methods aim to detect and track groups of people, since the members of each group usually exhibit the same motion pattern and in certain scenarios (e.g., high-density crowds) tracking individuals might not be feasible.

2.1. Tracking and detecting individuals. Rodriguez *et al*. (2011) proposed an algorithm in which crowd motion patterns are learned as priors from a large database of crowd videos collected from internet. The purpose is to enhance tracking individuals by using these motion patterns.

Tang *et al*. (2013) argue that the best performance can be achieved by training people detectors on failure cases. A joint people detector is proposed that is based on detection of both single persons and pairs of people, exploiting common patterns of occlusions that are considered the failure cases.

Heili and Odobez (2013) introduced a detection-based multi-target tracking method, where the tracking task is formulated as a statistical labeling process. The problem is solved using the *conditional random field* model, which depends on detection pairs to model pairwise features and color similarities/dissimilarities.

Wen *et al*. (2016) proposed a tracking algorithm that utilizes dense structures on affinity hypergraphs. Affinity was measured using appearance, motion and trajectory smoothness cues. Data association was improved by considering temporal similarities of tracklets. Unfortunately, the applicability of such methods is usually limited because of high complexity.

The majority of multi-object tracking approaches focus on improving performances of detectors or developing better data association schemes. A

coarse-to-fine multi-object tracking algorithm was used by Gong *et al*. (2016). A faster R-CNN network was applied for detection, the Kalman filter was used for coarse tracking, and a local sparse optical flow model was exploited for fine tracking.

Another detection-based tracking method was proposed by Zhang *et al*. (2015), where detections are linked into tracklets, and separate tracklets are further linked to form longer trajectories. The trajectory models are re-learned automatically using a unified algorithm.

Yang and Nevatia (2014) presented an approach for multi-target tracking where detections are grouped into tracklets to form the resulting tracks. An online learned *conditional random field* model (solved by energy minimization) was used for tracking. To discriminate between targets, motion and appearance models as well as a set of pairwise functions were included into the energy function.

Hofmann *et al*. (2013) presented a unified hierarchical multi-object tracking approach that is formulated as a three-stage maximum *a posteriori* problem with different parameters at each stage.

Ren *et al*. (2018) proposed a tracking method that fused density maps with visual object trackers. In this method, a *sparse kernelized correlation filter* was introduced to suppress target variations caused by occlusions, illumination changes, and spurious responses. Then, the people tracker fuses the filter response map with a crowd density map using a convolutional neural network.

2.2. Tracking and detecting groups. Edman *et al*. (2013) presented a method where groups are detected using integral channel features, and tracked using a Gaussian mixture of probabilistic density filters. The results are transformed from image coordinates into the ground-plane coordinates so that groups can be defined in terms of distances and speed differences.

A combination of low-level keypoint tracking, mid-level patch tracking, and high-level group evolution was proposed by Zhu *et al.* (2014; 2018). Keypoint tracking was used to obtain local motions at the low level, while mid-level patches were used to represent the crowd and then tracked through appearance variations. In this approach, a KLT tracker was used, and then the KNN-based clustering algorithm was utilized. The patches were organized into a hierarchical structure representing evolution of groups.

Ge *et al*. (2012) introduced a method for detecting and tracking small groups of people. A full-body HOG detector together with a correlation tracker were used for individuals. To discover groups, the Hausdorff distance was defined on pairwise proximities and velocities. A similar approach was presented by Raj and Poovendran (2014). Here, *reversible jump Markov chain Monte Carlo* was used to obtain overlapping rectangles that best cover the foreground pixels. However, such methods would probably fail in dense crowds with a high level of occlusions.

For detecting and tracking interacting groups in crowds, a framework based on the social force model was proposed by Mazzon *et al*. (2013). A buffered graph-based tracker was used to track the detected groups by linking the interaction centers of these groups. Interactions were predicted by an iterative minimization of the error between measurements and predictions.

Shao *et al*. (2014) introduced group profiling to understand the group-level dynamics and properties. A set of tracklets detected by KLT feature tracker was used to form a group. Visual descriptors were provided to quantify inter- and intra-group properties, which were used for crowd classification and analysis.

Based on the notion that groups move in the same direction, while avoiding people moving in the opposite direction, Zhang *et al*. (2018) proposed a crowd motion segmentation method. The method is primarily designed for specific scenarios of behavioral analysis, which may limit its use in other applications.

A joint individual-group tracking framework using particle filtering was proposed by Bazzani *et al*. (2012) for group tracking. In this work, the individual-group state space is factorized into two dependent subspaces where the joint individual-group distributions are shared between individuals and groups. Thus, relations between models of individuals and groups are established.

Yu *et al*. (2016) proposed a groupwise association and tracking method that was based on the individual group information and group correlations. A nonrigid 2D *Thin-Plate transform* was used to model the associations within a group, and then shrinking, growing, and merging operations were uses to refine the composition of each group.

To quantify and detect collective motions, Wang *et al*. (2017) introduced a framework that investigates the time-varying dynamics of individuals to find similarities in their motions, while a more comprehensive understanding of the crowd is obtained by a multi-stage clustering. Then, multi-view clustering was proposed by Wang *et al.* (2020a) for coherent groups detection. In this work, individuals are represented by a structural descriptor and clustered by a a self-weighted multi-view clustering. Additionally, a group detector is used to count the number of groups.

It should be noted, however, that no specific definitions of groups are used in these works. This notion is, in general, vague and often ill-defined, and we also use the concept of groups in its basic, informal sense.

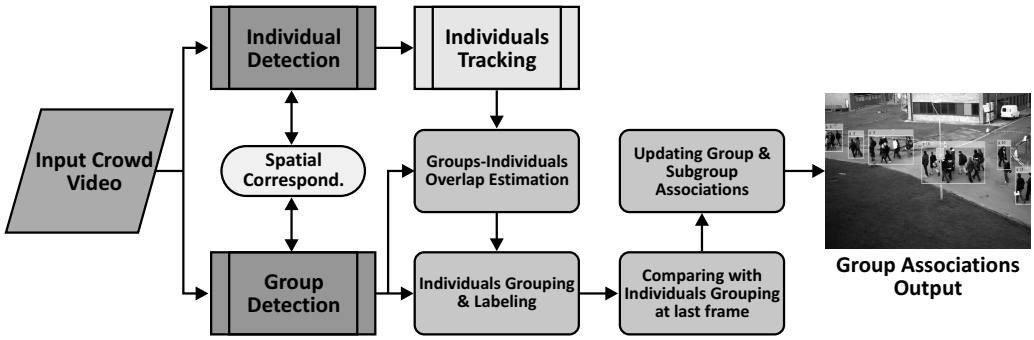


Fig. 1. Block diagram of building group associations from tracking individuals (Scheme 1).

3. Proposed model

Crowd analysis, e.g., classification of typical crowd behaviors (Zitouni *et al*., 2019b), usually requires data related to both individuals and groups. Although individuals and groups can be separately detected and tracked, simultaneous analysis better reflects interactions between crowd components and facilitates further analysis.

However, to the best of our knowledge, no general models unifying data association of individuals and groups in crowd analysis seem to exist, especially for data acquired with a high level of uncertainty (limited performances of detectors and trackers). Therefore, we propose two non-deterministic schemes representing two opposite scenarios of data association in crowd analysis. In practice, the schemes can be adaptively combined, depending on the actual characteristics of processed video-sequences.

Both schemes use detectors of individuals and groups, but trackers are applied selectively. The first scheme is based on tracking individuals. Then, the individual tracks (in conjunction with data on membership of individuals in groups) are employed to perform group data association (labeling groups for monitoring their evolution). In the second scheme, the approach is reversed, i.e., only groups are tracked and group tracking results are used to perform data association for individuals (based on their group memberships).

In both schemes, the ultimate results specify changes within the crowd, including (dis)appearance of individuals and groups, evolution of groups (e.g., splitting, merging) and relations between groups and individuals (joining or leaving). Figure 1 shows a block diagram illustrating the first scheme, while Fig. 2 depicts a block diagram for the second scheme.

To emphasize independence of the schemes from the raw-data tools (detectors and trackers), exemplary tools used in the performed experiments are discussed at the end of this section.

3.1. Notation and definitions. In both schemes, we use the specific notation to represent detection and tracking results (for individuals and groups) in the current video-frame t and, if applicable, in previous frames $\{t-1, t-2, \ldots, t-K, \ldots\}$. Thus

$$I(t) = [i_1(t), i_2(t), \ldots, i_{N_t}(t)] \tag{1}$$

is the vector representing N_t people identified in the current frame t by detector(s) of individuals. Similarly

$$G(t) = [g_1(t), g_2(t), \ldots, g_{M_t}(t)] \tag{2}$$

is the vector representing M_t groups extracted in the current frame t by detector(s) of groups.

To indicate relations between groups and individuals, we introduce the notation

$$\widehat{gi}_{m,n}(t) \equiv \text{est}[i_n(t) \in g_m(t)] \tag{3}$$

to numerically estimate (from interval $[0, 1]$, e.g., using *probabilities*) the confidence level that in frame t the individual $i_n(t)$ is a member of group $g_m(t)$.

Correspondingly,

$$\widehat{gi}_{m,n}(t-K, t) \equiv \text{est}[i_n(t) \in g_m(t-K)] \tag{4}$$

numerically estimates the confidence level that individual $i_n(t)$ from frame t was a member of group $g_m(t-K)$ in frame $t-K$. In the same manner,

$$\widehat{ii}_{m,n}(t-K, t) \equiv \text{est}[i_n(t) = i_m(t-K)] \tag{5}$$

is the numerical estimate that the individual $i_n(t)$ from frame t is the same as individual $i_m(t-K)$ from frame $t-K$.

Finally,

$$\widehat{gg}_{m,n}(t-K, t) \equiv \text{est}[g_n(t) = g_m(t-K)] \tag{6}$$

is the numerical estimate that group $g_n(t)$ in frame t is the same as group $g_m(t-K)$ from frame $t-K$.

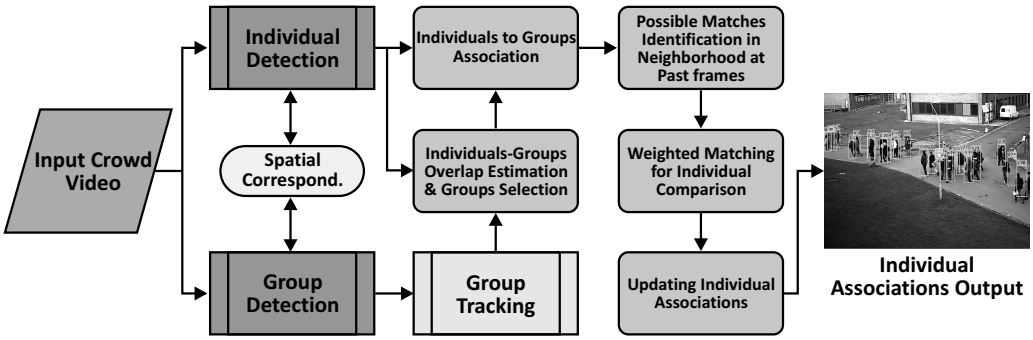


Fig. 2. Block diagram of building individual associations from group tracking (Scheme 2).

Based on Eqns. (1)–(3), we estimate in frame t the levels of individual memberships in groups by an $M_t \times N_t$ matrix

$$\widehat{GI}(t) = \begin{bmatrix} \widehat{gi}_{1,1}(t) & \dots & \widehat{gi}_{1,N_t}(t) \\ \widehat{gi}_{2,1}(t) & \dots & \widehat{gi}_{2,N_t}(t) \\ \dots & \dots & \dots \\ \widehat{gi}_{M_t,1}(t) & \dots & \widehat{gi}_{M_t,N_t}(t) \end{bmatrix}. \tag{7}$$

Zero columns of $\widehat{GI}(t)$ indicate individuals not belonging to any detected group, while zero rows represent groups with no members. Usually, such cases represent false-positive detections which should be removed, i.e., the numbers N_t or M_t are appropriately reduced.

It should be highlighted that, in general, the values of $\widehat{gi}_{m,n}(t)$ are not equivalent to probabilities. Usually, they do not sum up to 1 along columns. Each number is estimated separately and often not related to other values in the same column. In particular, it is possible to have multiple ones in a column, if a group is a sub-group of another group, or if the same group is independently identified by a different detector. More details on typical methods used to obtain values $\widehat{gi}_{m,n}(t)$ from actual visual data are given in the following subsection.

3.1.1. Estimating $\widehat{gi}_{m,n}$. In the simplest scenario, the values of $\widehat{gi}_{m,n}(t)$ can be estimated by the ratio

$$\widehat{gi}_{m,n}(t) = \frac{\text{area}(BB_{i_n} \cap BB_{g_m})}{\text{area}(BB_{i_n})}, \tag{8}$$

where BB_{i_n} is the bounding box of the n-th person and BB_{g_m} is the outline of the m-th group.

Alternatively, objects (including individuals and groups) can be represented by *density functions* (e.g., Li *et al.*, 2018) or by *heat maps* (e.g., Zhou *et al.*, 2019) where object outlines are defined by (near-)zero values of these functions. Then Eqn. (8) can be generalized to

$$\widehat{gi}_{m,n}(t) = \frac{\int f_{i_n}(x,y) f_{g_m}(x,y)\,\mathrm{d}x\,\mathrm{d}y}{\int f_{i_n}(x,y)\,\mathrm{d}x\,\mathrm{d}y} \tag{9}$$

where f_{i_n} and f_{g_m} are, respectively, the density functions (or heat maps) of the n-th person and of the m-th the group in an image with $X \times Y$ coordinates.

Actually, one of the exemplary group detectors presented in Section 3.4 applies the probability density approach.

Remarkably, computational costs of estimates $\widehat{gi}_{m,n}$ are very low. It can be noted that both Eqns. (8) and (9) require just a single scan of a frame after it is processed by the detectors of individuals and groups. Thus, real-time performances are easily obtainable, provided that both detectors perform in real time.

3.2. Group associations from tracking individuals. In this scheme (referred to as Scheme 1) associations between groups of the current frame t and a past frame $t-K$ are inferred from tracking individuals. Usually, we consider the neighboring frames t and $t-1$ but, as shown below, any past frames can be alternatively used.

In both frames, groups and individuals are detected so that the matrices $\widehat{GI}(t)$ and $\widehat{GI}(t-K)$ (cf. Eqn. (7)) are available. We assume that N_{t-K} individuals and M_{t-K} groups occur in frame $t-K$, and N_t individuals and M_t groups exist in frame t.

First, from tracking individuals we build an $N_{t-K} \times N_t$ matrix $\widehat{II}(t-K,t)$ containing estimates (see Eqn. (5)) that individual $i_m(t-K)$ from frame $t-K$ is the same as individual $i_n(t)$ in the current frame t:

$$\widehat{II}(t-K,t) = \begin{bmatrix} \widehat{ii}_{1,1}(t-K,t) & \dots & \widehat{ii}_{1,N_t}(t-K,t) \\ \widehat{ii}_{2,1}(t-1,t) & \dots & \widehat{ii}_{1,N_t}(t-K,t) \\ \vdots & \vdots & \vdots \\ \widehat{ii}_{N_{t-K},1}(t-K,t) & \dots & \widehat{ii}_{N_{t-K},N_t}(t-K,t) \end{bmatrix}. \tag{10}$$

In general, matrices $\widehat{II}(t-K,t)$ should satisfy

a straightforward requirement that sums along its rows (columns) must not exceed *one* since an individual in the current (past) frame cannot physically correspond to more than one person from the past (current) frame. In particular, *zero-valued* columns (or rows) are possible; they represent newly appearing crowd members (or disappearing crowd members).

Afterwards, we can estimate memberships of the current-frame individuals in the past-frame groups. Let $\widehat{GI}(t-K,t)$ be the matrix of estimates (e.g., probabilities) that individuals from frame t were members of groups in frame $t-K$:

$$\widehat{GI}(t-K,t) = \begin{bmatrix} \widehat{gi}_{1,1}(t-K,t) & \dots & \widehat{gi}_{1,N_t}(t-K,t) \\ \widehat{gi}_{2,1}(t-K,t) & \dots & \widehat{gi}_{2,N_t}(t-K,t) \\ \vdots & \vdots & \vdots \\ \widehat{gi}_{M_{t-K},1}(t) & \dots & \widehat{gi}_{M_{t-K},N_t}(t-K,t) \end{bmatrix}. \tag{11}$$

This matrix can be estimated as

$$\widehat{GI}(t-K,t) = \widehat{GI}(t-K,t-K) \times \widehat{II}(t-K,t) \tag{12}$$

and its size is $M_{t-K} \times N_t$.

Given such a history of the individual memberships (which can span, through more matrix multiplications, over any number or frames) we are able to trace correspondences between groups, i.e., data associations between groups in the current frame t and groups in a past frame $t-K$ can be established. This is modeled by $\widehat{GG}(t-K,t)$ matrix of size $M_{t-K} \times M_t$ obtained as

$$\widehat{GG}(t-K,t) = \widehat{GI}(t-K,t) \times \widehat{GI}(t,t)^T. \tag{13}$$

The elements $\widehat{gg}_{m,n}(t-K,t)$ of this matrix are, formally, non-deterministic estimates of the numbers of individuals shared by group $g_m(t-K)$ from frame $t-K$ and group $g_n(t)$ from frame t.

Consequently, groups can be deterministically associated (based on the maximum estimated number of shared members) at any time instances when such an association is needed.

The label of a past-frame group $g_m(t-K)$ is (forward) linked to the label of a current-frame group $g_n(t)$ corresponding to the maximum value of in the m-th column of matrix $\widehat{GG}(t-K,t)$. Similarly, the label of a current-frame group $g_n(t)$ is (backward) linked to the label of a past-frame group $g_m(t-K)$ corresponding to the maximum value of in n-th row of the matrix $\widehat{GG}(t-K,t)$.

Because links are created bidirectionally, multiple label associations are possible. For example, the same current group can be backward linked to several past groups (which indicates group merging) or the same past group can be forward linked to multiple current groups (which represents group splitting). In this way, labels flexibly propagate through time and informative details on the crowd evolution in the monitored scene are provided.

It should be noted that *all-zero* columns of matrix $\widehat{GG}(t-K,t)$ represent groups that disappear and their labels should be terminated. Correspondingly, *all-zero* rows of $\widehat{GG}(t-K,t)$ indicate newly appearing groups which should be assigned new labels.

3.3. Individual associations from tracking groups.

In this scheme (referred to as Scheme 2), individual associations between the current frame t and a past frame $t-K$ are inferred from tracking groups (to which people belong), i.e., only the group tracker is used. Again, detectors of groups and individuals are applied so that the matrices $\widehat{GI}(t)$ and $\widehat{GI}(t-K)$ (cf. Eqn. (7)) are available, where N_{t-K}, N_t, M_{t-K} and M_t indicate, respectively, the numbers of detected individuals and groups in both frames.

The associations between groups in the current frame t and the past frame $t-K$ are obtained from tracking results. Formally, these associations are represented by a $M_{t-K} \times M_t$ matrix $\widehat{GG}(t-K,t)$ containing estimates (cf. Eqn. (6)) that group $g_m(t-K)$ from the past frame $t-K$ is the same as group $g_n(t)$ in the current frame t):

$$\widehat{GG}(t-K,t) = \begin{bmatrix} \widehat{gg}_{1,1}(t-K,t) & \dots & \widehat{gg}_{1,M_t}(t-K,t) \\ \widehat{gg}_{2,1}(t-K,t) & \dots & \widehat{gg}_{2,M_t}(t-K,t) \\ \vdots & \vdots & \vdots \\ \widehat{gg}_{M_{t-K},1}(t-K,t) & \dots & \widehat{gg}_{M_{t-K},M_t}(t-1,t) \end{bmatrix}, \tag{14}$$

where the estimates $\widehat{gg}_{m,n}(t-K,t)$ are provided by the group tracker. In our experiments, a group tracker derived from the Kalman filter (see Section 3.4) is applied as the main baseline tracker, but any alternative tracker can be used instead.

Now, the objective is to identify the most credible associations between individuals from frames $t-K$ and t, without explicitly tracking them.

We estimate associations between the i-th individual from frame $t-K$ and the j-th individual from frame t (they are identified by subscripts (i,j)) using the

following matrix:

$$\widehat{II}_{(i,j)}(t-K,t)=\begin{bmatrix} \widehat{ii}_{(i,j)1,1}(t-K,t) & \dots \\ \vdots & \vdots \\ \widehat{ii}_{(i,j)M_{t-K},1}(t-K,t) & \dots \\ \end{bmatrix}\begin{matrix} \widehat{ii}_{(i,j)1,M_t}(t-K,t) \\ \vdots \\ \widehat{ii}_{(i,j)M_{t-K},M_t}(t-K,t) \end{matrix}\Bigg], \tag{15}$$

where

$$\widehat{ii}_{(i,j)m,n}(t-K,t)=\widehat{gi}_{m,i}(t-K)\cdot\widehat{gi}_{n,j}(t)\cdot\widehat{gg}_{m,n}(t-K,t) \tag{16}$$

i.e., we combine estimates of individual memberships in groups (obtained from detectors) with group associations $\widehat{gg}_{m,n}(t-K,t)$ provided by the group tracker.

Altogether, associations between all individuals from frames $t-K$ and t can be represented by a 4D tensor (a 2D matrix of 2D matrices):

$$\widehat{II}(t-K,t)=\begin{bmatrix} \widehat{II}_{(1,1)}(t-K,t) & \dots \\ \vdots & \vdots \\ \widehat{II}_{(N_{t-K},1)}(t-K,t) & \dots \\ \end{bmatrix}\begin{matrix} \widehat{II}_{(1,N_t)}(t-K,t) \\ \vdots \\ \widehat{II}_{(N_{t-K},N_t)}(t-K,t) \end{matrix}\Bigg]. \tag{17}$$

The above formalism of individual associations from group tracking does not exploit the visual appearances of crowd members. Thus, it is equally applicable to crowds of distinctively different people and crowds of similarly looking (e.g., uniformed) individuals. In fact, visual appearances (if sufficiently diversified and obtainable from acquired frames) can further improve the credibility of the associations. In the implemented feasibility study, we actually employed this supplemental option in a simple, yet often effective, way.

We just consider the outline (bounding box) of a detected individual, a key point (key region) for which any key point descriptor can be calculated. As an example, we use SURF descriptors (in the RGB space). In the case of the same people (with approximately the same sections of their bodies outlined in both frames) similar descriptor values are expected. This can be used to up-value associations provided by Eqn. (17).

For practical reasons, we primarily match descriptors of individuals within groups with non-zero estimates $\widehat{gg}_{m,n}(t-K,t)$. However, individuals from other groups are also matched (but only for limited-size geometric neighborhoods). In other words, we assume that a person may change a group, but cannot move too far (even if running) within a short period of time. Since the fastest humans can only run at a speed of 10 m/s, the position differences between subsequent frames cannot exceed 0.3 to 0.4 m (for typical fps rates) so that the suitable size of the neighborhood can be easily estimated for a given setup (fps rate, distance from the monitored scene, etc.) of the surveillance camera, taking into account the most extreme scenarios expected.

In general, tensors $\widehat{II}(t-K,t)$ are very sparse with only a few non-zero elements, which simplifies data associations between individuals. The deterministic association (i.e., the label continuity) for the j-th individual from frame t is defined by the maximum element $\widehat{ii}_{(i,j)m,n}(t-K,t)$ in tensor $\widehat{II}(t-K,t)$. However, since multiple deterministic associations are not possible for people, if the i_1-th maximum-score association for the j-th individual is already taken by another person k (with a higher value of $\widehat{ii}_{(i_1,k)m,n}(t-K,t)$), the second best choice is selected, etc.

If for the j-th individual from frame t, all values $\widehat{ii}_{(i,j)m,n}(t-K,t)$ are *zeros* (or if all non-zero values are "taken" by other crowd members) a new label will be assigned for that individual. If, correspondingly, the same happens for the i-th individual from frame $t-K$, its label is removed from the list of active crowd members.

Unfortunately, in real-world surveillance misdetection of people frequently happens. Therefore, we recommend to perform data associations between the current-frame individuals and individuals across a few most recent frames. Then, the values of tensor $\widehat{II}(t-K,t)$ should be weighed by factors w_K decreasing with the temporal distance between the frames, e.g., $w_1=1$, $w_2=0.9$, $w_3=0.8$, etc. Eventually, the individuals from the current frame t would be linked to individuals from any of the most recent frames, based on the highest scores.

3.4. Exemplary detectors and trackers. Our model can use any detectors of individuals and groups (which generate outlines, e.g., bounding boxes, of individual silhouettes or group shapes). In exemplary feasibility study implementations, we employed two types of people detectors and two group detectors. In both cases, one example is a classical detector, while the other is CNN-based. The classical detector of individuals is the aggregated channel features (ACF) detector (Dollár *et al*., 2014), which was originally trained on an INRIA person data set (Dalal and Triggs, 2005)).

Classical group detection is performed by motion-based segmentation of the scene foreground. The Gaussian mixture model of dynamic textures (GMM-of-DT) from Zitouni *et al*. (2016) is used to identify motion saliency (from a sequence of frames).

The salient motion is modeled using a GMM framework with K Gaussian distributions. The probability that a certain pixel represents a group is defined in a standard way as

$$p(y_t) = \sum_{i=1}^{K} w_i \left(\frac{1}{|2\pi\Sigma_k|^{1/2}} e^{-1/2(y_t-\mu_k)^T \Sigma_k^{-1}(y_t-\mu_k)} \right), \tag{18}$$

where w_i is the weight of the k-th Gaussian component and μ_k and Σ_k are its mean and covariance, respectively. A threshold value is then applied to approximate the group regions, and morphological analysis is performed to smooth their shapes. Note that this detector employs data from a number of consecutive frames.

As an alternative, CNN-based detectors (originally presented by Zitouni *et al*. (2019a)) have been used. In this method, regions are proposed in crowd images and classified into individuals, small groups and large groups. This detector uses only data from a single frame.

Exemplary trackers of individuals and groups are as below.

The tracker of individuals (which is used to obtain elements of matrices $\widehat{II}(t-K,t)$, cf. Eqn. (10), is based on the publicly available implementation of the JPDA tracking method by Rezatofighi *et al*. (2015). In this algorithm, the states of N_t individuals and their R_t measurements are respectively denoted by $\{x_1(t), \ldots, x_{N_t}(t)\}$ and $\{z_1(t), \ldots, I_{R_t}(t)\}$.

The state $x_n(t)$ of an individual combines his/her position and velocity. The detected positions (which are usually noisy and cluttered) are incorporated into measurements $z_r(t)$. Then, using a linear Gaussian model, the tracking probabilities are obtained as

$$p(d_r^n(t)) \propto \begin{cases} (1-p(i_n))\beta, & r=0, \\ p(i_n)\cdot\mathcal{N}(z_r(t), \hat{i}_n(t), \Sigma_S), & r\neq 0. \end{cases} \tag{19}$$

where $p(i_n)$ is the detection probability of individual $i_n(t)$ estimated by the ACF detector, β is the false detection density and $\hat{i}_n(t)$ is the predicted position of individual $i_n(t)$ in frame t, and Σ_S is the covariance matrix of the Kalman filter. Note that $\mathcal{N}$ indicates the normal distribution.

The problem is then reformulated as an integer linear program, and solved to achieve the maximum likelihood. We use these results as estimates of matrices $\widehat{II}(t-K,t)$, see Eqn. (10), i.e., $\widehat{ii}_{m,n}(t-K,t) = p(d_r^n(t-K))$.

The Kalman filter is also applied as the baseline tracker for groups, i.e., the tracking estimates $\widehat{gg}_{m,n}(t-K,t)$ in Eqns. (6) and (14) are obtained as

$$\widehat{gg}_{m,n}(t-K,t) = p(x_t|z_{1:t-K}), \tag{20}$$

where x_t is the state (predicted position) of the group (based on the evolution of the prior state at $t-K$) and z is the Kalman observation.

4. Exemplary experimental results

To evaluate the practicality and usefulness of the proposed model, both schemes were tested on a number of videos showing diversified crowd behaviors. Unfortunately, even the most recent data sets on crowd structure analysis, (e.g., Wang *et al.*, 2020c; 2020b) focus on performance evaluation in various factors related to detection and tracking.

In general, we do not intend to test performances of any particular detectors/trackers or to improve trackers. Nevertheless, we found that our methodology can enhance trackers by removing falsely initiated tracks and terminating tracks incorrectly continued (belonging to people who disappeared from the scene) as highlighted in Section 4.2.

Because of the above assumptions, the test video-sequences have been chosen from PETS (Ferryman and Shahrokni, 2009), Parking Lot (Dehghan *et al*., 2015) and Town Center (Benfold and Reid, 2011) collections. These popular data sets (even if not particularly challenging for testing trackers and/or detectors) contain highly diversified crowd behaviors, and are sufficiently complex for evaluating the proposed model. In some tests, no suitable benchmarks exist and we had to identify (sometimes subjectively) ground-truth reference results for comparison.

We focused on six videos, four from the PETS data set, i.e., $S2\_L2\_14-55$, $S1\_L1\_13-57$, $S2\_L1\_12-34$, and $S1\_L2\_14-06$ sequences, one from the *Parking Lot* data set and one from the *Town Center* data set. These videos contain diversified crowd densities, distributions of individuals and groups, motion patterns, occlusions and visibility conditions. Exemplary frames from the videos (with some detection results) are given in Fig. 3.

In all tests, detectors and trackers mentioned in Section 3.4 are used as the baseline algorithms. Individuals and groups are represented by labeled bounding boxes, where the same labels indicate continuity from the past frames. Labeling is driven either by tracking results (if our model is not applied), or by the scores produced by our model (as explained in Sections 3.2 and 3.3). Since individuals are assigned to groups (see Eqn. (7)) group labels are linked to labels of group members. Such data are essential, in particular, for the interpretation of the crowd structure and evolution.

4.1. Evolution of crowd structure. These tests represent the most significant intended application of the proposed framework. First, we demonstrate how our model helps to improve group evolution analysis (formation and disappearance of groups, changes in group memberships, etc.).

Such results can be subsequently used to identify crowd behaviors present in the scene, e.g., using the

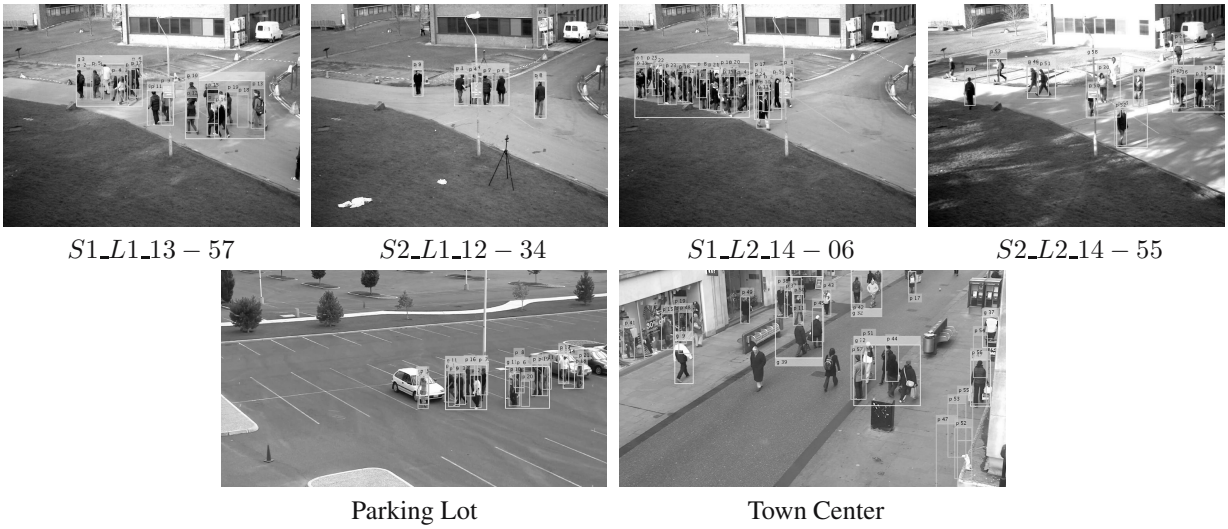


Fig. 3. Exemplary frames from tested video-sequences.

socio-cognitive categorization discussed by Zitouni *et al*. (2020), where four main categories are considered. The categories are: *individualistic*, *group* (with *leader-following* as a special case) and *social interaction* behaviors.

Plots in Fig. 4 visualize exemplary crowd evolutions for various behaviors. In $S1\_L1\_13-57$ and *Parking Lot* sequences, the crowd mainly consists of large groups with a few smaller groups temporarily formed. This is characteristic for *group* behaviors.

In $S2\_L1\_12-34$ sequence, the groups are small but (when formed) they last for a noticeable period of time. This indicates *social interaction* behavior where people incidentally interact with each other. In $S2\_L2\_14-55$ sequence, there is a large number of groups, mostly small, that quickly disappear after being formed. This indicates randomness of individual motions and general volatility of the crowd structure, which implies *individualistic* behavior.

To demonstrate that Scheme 1 actually provides improvements, its results are compared with the same analysis performed using only the baseline group tracking; the results are shown in Table 1. The performance criteria include the average lifespan of group labels, the average number of labels per frame, and the total number of group labels in the whole sequence.

In general, it is desirable to have lower numbers of group labels per frame. With the same detectors and trackers, this simply means more compact (and presumably less redundant) representation of crowds. Thus, our model is superior in all examples.

However, the total number of group labels and the average label lifespan should depend on the actual crowd behavior. In *group* behaviors (where groups exist for a significant period of time) longer lifespans and lower numbers of labels better reflect the reality (e.g., $S1\_L1\_13-57$ and *Parking Lot* sequences). For *individualistic* behaviors (e.g. $S2\_L2\_14-55$ sequence) the total number of labels should be rather high with very short lifespans (because groups actually do not exist; they are only instantaneously formed and quickly disappear). Again, our model provides more realistic estimates.

For *social interaction* behaviors, the total number of group labels is rather unpredictable, but the groups (usually small) should last for a while. For the exemplary sequence $S2\_L1\_34-24$ the results of our model are, therefore, comparable (if not superior) to the tracking-only baseline.

In Fig. 5, we show selected results by Scheme 2 where group switching by crowd members is presented. Two sequences from Fig. 4 are used and (for the sake of figure readability) only some crowd members are displayed. In $S1\_L1\_13-57$ sequence (where two large groups dominate) the majority of individual labels stably remain in the same group. In $S2\_L1\_12-34$ sequence (with many small groups occasionally merging and splitting) more frequent group switching can be noticed for many crowd members.

4.2. Other prospective applications. Although analysis of crowd evolution is the main intended application of our model, we have also identified that the model can be used as a supporting tool to improve tracking results for individuals. We tested this approach partly because the corresponding benchmark references on the PETS data set are available.

First, we compared the JPDA tracker (which is one of our baseline trackers for individuals) with the same

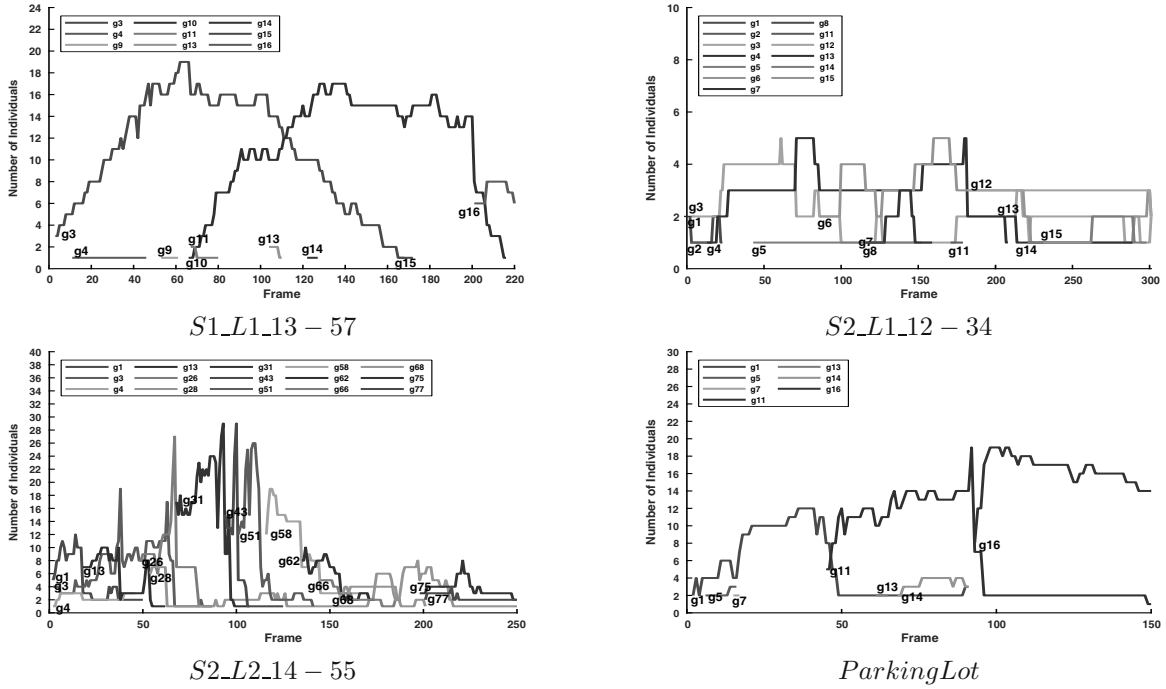


Fig. 4. Plots of the size and lifespan of groups over sequences of frames (Scheme 1 used). Groups with very short lifespans are ignored.

Table 1. Comparison between results of Scheme 1 (ours) and the tracking-only baseline (Bs) in group evolution analysis.

Sequence	Total no. group labels		Avg. group label lifespan (frames)		Avg. no of group labels per frame		Behavior type
	Ours	Bs	Ours	Bs	Ours	Bs	
$S1\_L1\_13-57$	16	20	113.12	73.25	1.93	5.70	Group
$S2\_L1\_12-34$	15	18	117.50	102.25	4.12	5.14	Social Interaction
$S2\_L2\_14-55$	88	39	30.21	52.76	5.56	16.28	Individualistic
Parking Lot	16	17	64.4	44.1	2.08	4.65	Group

tracker supported by our model. For a fair comparison, we used the original script and ground-truth provided by Rezatofighi *et al*. (2015) and the benchmark results were taken from the literature.

JPDA tracking data are applied to create group associations by Scheme 1, and the associations are subsequently used to modify the tracker parameters.

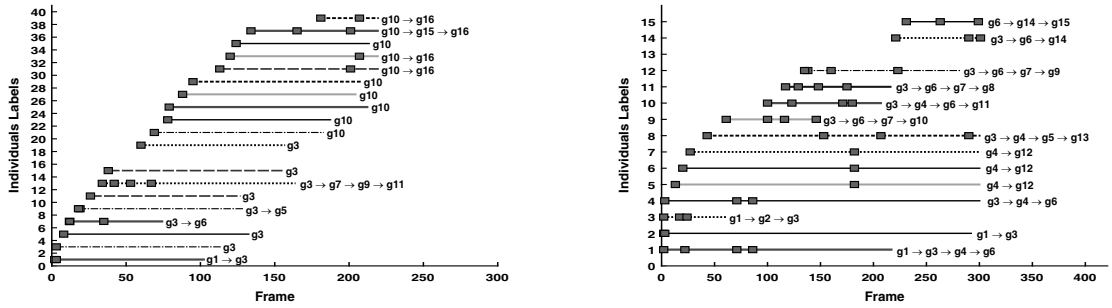
Table 2 compares our results with the original JPDA tracker using various measures, i.e., precision, recall, F-measure and multiple object tracking accuracy (MOTA). The proposed approach outperforms the JPDA tracker in terms of recall and F-measure for all test sequences, while providing better or comparable results in other metrics. We expect similar improvements for any other baseline tracker of individuals.

Finally, we attempted to replace actual trackers of individuals by Scheme 2, where only group tracking is available (see Section 3.3) and no actual tracking is performed on individuals (which makes this test very challenging). Nevertheless, the results in Table 3 show that on exemplary (and, actually, on many more) PETS sequences Scheme 2 is comparable (and sometimes superior) to popular trackers of pedestrians in terms of precision, recall and F-measure.

5. Conclusions

The paper presents a framework of two alternative schemes (including their non-deterministic definitions, exemplary embodiments and experimental feasibility studies) for data association in visual surveillance of crowd behavior. By using inter-dependencies between the results of the baseline detectors and trackers of individuals and groups, we provide a mathematical model for maintaining data association continuity in scenarios where results of tracking/detection algorithms are temporarily corrupted or disrupted (due to visual conditions, occlusions, bad weather, etc.).

The schemes can be integrated with any combination of detection and tracking techniques, including CNN-based detectors/trackers, as long as they provide



$S1\_L1\_13-57$ $S2\_L1\_12-34$

Fig. 5. Plots of group switching patterns by selected crowd members, based on results with Scheme 2. The lines terminate when the labels are discarded.

Table 2. Performances of tracking individuals. In the proposed method, Scheme 1 is incorporated into the JPDA tracker.

Sequence	Method	Precision	Recall	F-measure	MOTA
$S1\_L1\_13-57$	Proposed	**91.2**	**69.6**	**78.9**	**61.4**
	Rezatofighi *et al*. (2015)	86.2	58.3	69.6	48.2
$S2\_L1\_12-34$	Proposed	80.6	**98.8**	**88.8**	73.4
	Rezatofighi *et al*. (2015)	**83.9**	92.9	88.2	**74.9**
$S1\_L2\_14-06$	Proposed	84.1	**55.5**	**66.9**	**42.7**
	Rezatofighi *et al*. (2015)	**86.8**	40.3	55.0	32.8
$S2\_L2\_14-55$	Proposed	**88.7**	**75.1**	**81.3**	52.4
	Rezatofighi *et al*. (2015)	85.6	**75.1**	80.0	**60.9**

outlines (e.g., bounding boxes) of the detected/tracked groups and individuals. Exemplary combinations of the selected *state-of-the-art* detection/tracking algorithms are used in the presented feasibility-study implementations.

The feasibility tests were performed on exemplary video-sequences (from the PETS data set and two other publicly available data sets) and the proposed schemes have shown their advantages. In particular, the produced data associations (which usually, but not exclusively, represent fluctuations in group numbers, sizes, and individual memberships in groups), provide important clues for the subsequent analysis of the crowd evolution, including socio-cognitive categorization of the crowd behavior, in the monitored environments.

Because no benchmark results are available for the tests targeting the main area of intended applications, we have built our own references (based on results provided by the baseline trackers) in these major tests. Supplementary tests were performed in pure tracking tasks (for which the proposed schemes are not intended) because the availability of the *state-of-the-art* benchmark results. In general, it was found that by using the proposed model:

- The evolution of crowd structure can be monitored more reliably then by using only detectors and trackers of groups and individuals.
- Even in standard tasks of tracking people, the model (though not intended for such applications) provides results superior/comparable to the benchmarks from publicly available trackers of individuals.

Computationally, the model is based primarily on matrix operations (where the sizes of matrices correspond to the numbers of detected individuals and groups), i.e., it can be applied in real-time problems, including scenarios with large crowds of complex structures.

Two schemes proposed in the model are not exclusive (since both use the same baseline detectors of groups and individuals). Actually, it is recommended to use them in parallel to further improve performances in various problems of crowd analysis. Additionally, the parallel involvement of multiple detectors and/or trackers is also suggested in the model.

In the future works, we intend to adapt the proposed model to multi-camera setups (including mobile cameras) where the additional factor of spatial data associations (switching between cameras) should be combined with the temporal data associations.

Acknowledgment

Financial support of Khalifa University is gratefully acknowledged.

References

Bazzani, L., Cristani, M. and Murino, V. (2012). Decentralized particle filter for joint individual-group tracking,

Table 3. Data associations for individuals by Scheme 2 compared to specialized trackers of pedestrians.

Sequence	Method	Precision	Recall	F-measure
$S1\_L2\_14-06$	**Scheme 2**	90.4	**40.4**	**55.8**
	Rezatofighi *et al*. (2015)	86.8	40.3	55.0
	Milan *et al*. (2014)	86.4	38.5	53.3
	Berclaz *et al*. (2011)	92.6	21.4	34.8
$S2\_L2\_14-55$	**Scheme 2**	**93.7**	57.3	77.1
	Rezatofighi *et al*. (2015)	85.6	75.1	80.0
	Milan *et al*. (2014)	89.8	65.5	75.7
	Wen *et al*. (2016)	90.3	71.2	79.6

IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, pp. 1886–1893, DOI: 10.1109/CVPR.2012.6247888.

Benfold, B. and Reid, I. (2011). Stable multi-target tracking in real-time surveillance video, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, USA*, pp. 3457–3464, DOI: 10.1109/CVPR.2011.5995667.

Berclaz, J., Fleuret, F., Turetken, E. and Fua, P. (2011). Multiple object tracking using k-shortest paths optimization, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**(9): 1806–1819, DOI: 10.1109/TPAMI.2011.21.

Bochinski, E., Senst, T. and Sikora, T. (2018). Extending IOU based multi-object tracking by visual information, *15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand*, pp. 1–6, DOI: 10.1109/AVSS.2018.8639144.

Ciaparrone, G., Luque Sanchez, F., Tabik, S., Troiano, L., Tagliaferri, R. and Herrera, F. (2020). Deep learning in video multi-object tracking: A survey, *Neurocomputing* **381**: 61–88, DOI: 10.1016/j.neucom.2019.11.023.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection, *IEEE Conference on Computer Vision and Pattern Recognition, San Diego, USA*, Vol. 1, pp. 886–893, DOI: 10.1109/CVPR.2005.177.

Dehghan, A., Modiri Assari, S. and Shah, M. (2015). GMMCP tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4091–4099, DOI: 10.1109/CVPR.2015.7299036.

Dollár, P., Appel, R., Belongie, S. and Perona, P. (2014). Fast feature pyramids for object detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**(8): 1532–1545, DOI: 10.1109/TPAMI.2014.2300479.

Edman, V., Andersson, M., Granström, K. and Gustafsson, F. (2013). Pedestrian group tracking using the GM-PHD filter, *European Signal Processing Conference (EUSIPCO), Marrakech, Morocco*, pp. 1–5.

Ferryman, J. and Shahrokni, A. (2009). PETS2009: Dataset and challenge, *12th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Snowbird, USA*, DOI: 10.1109/PETS-WINTER.2009.5399556.

Garcia-Martin, A., Sanchez-Matilla, R. and Martinez, J.M. (2017). Hierarchical detection of persons in groups, *Signal, Image and Video Processing* **11**(7): 1181–1188, DOI: 10.1007/s11760-017-1073-z.

Ge, W., Collins, R.T. and Ruback, R.B. (2012). Vision-based analysis of small groups in pedestrian crowds, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34**(5): 1003–1016, DOI: 10.1109/TPAMI.2011.176.

Gong, S., Han, H., Shan, S. and Chen, X. (2016). Actions recognition in crowd based on coarse-to-fine multi-object tracking, *Asian Conference on Computer Vision, Taipei, Taiwan*, pp. 478–490, DOI: 10.1007/978-3-319-54526-4_35.

Heili, A. and Odobez, J.-M. (2013). Parameter estimation and contextual adaptation for a multi-object tracking CRF model, *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS), Clearwater, USA*, pp. 14–21, DOI: 10.1109/PETS.2013.6523790.

Hofmann, M., Haag, M. and Rigoll, G. (2013). Unified hierarchical multi-object tracking using global data association, *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS), Clearwater, USA*, pp. 22–28, DOI: 10.1109/PETS.2013.6523791.

Jacques, J.C.S., Braun, A., Soldera, J., Musse, S.R. and Jung, C.R. (2007). Understanding people motion in video sequences using Voronoi diagrams, *Pattern Analysis and Applications* **10**(4): 321–332, DOI: 10.1007/s10044-007-0070-1.

Kasprzak, W., Wilkowski, A. and Czapnik, K. (2012). Hand gesture recognition based on free-form contours and probabilistic inference, *International Journal of Applied Mathematics and Computer Science* **22**(2): 437–448, DOI: 10.2478/v10006-012-0033-6.

Li, D., Zhu, J., Xu, B., Lu, M. and Li, M. (2018). An ant-based filtering random-finite-set approach to simultaneous localization and mapping, *International Journal of Applied Mathematics and Computer Science* **28**(3): 505–519, DOI: 10.2478/amcs-2018-0039.

Mazzon, R., Poiesi, F. and Cavallaro, A. (2013). Detection and tracking of groups in crowd, *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Krakow, Poland*, pp. 202–207, DOI: 10.1109/AVSS.2013.6636640.

Milan, A., Roth, S. and Schindler, K. (2014). Continuous energy minimization for multitarget tracking, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**(1): 58–72, DOI: 10.1109/TPAMI.2013.103.

Park, M.-W. and Brilakis, I. (2016). Continuous localization of construction workers via integration of detection and tracking, *Automation in Construction* **72**(Part 2): 129–142, DOI: 10.1016/j.autcon.2016.08.039.

Raj, K.S. and Poovendran, R. (2014). Pedestrian detection and tracking through hierarchical clustering, *International Conference on Information Communication and Embedded Systems, Chennai, India*, pp. 1–4, DOI: 10.1109/ICICES.2014.7033991.

Ren, W., Kang, D., Tang, Y. and Chan, A.B. (2018). Fusing crowd density maps and visual object trackers for people tracking in crowd scenes, *IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA*, pp. 5353–5362, DOI: 10.1109/CVPR.2018.00561.

Rezatofighi, S.H., Milan, A., Zhang, Z., Shi, Q., Dick, A. and Reid, I. (2015). Joint probabilistic data association revisited, *IEEE International Conference on Computer Vision (ICCV), Santiago, Chile*, pp. 3047–3055, DOI: 10.1109/ICCV.2015.349.

Rodriguez, M., Sivic, J., Laptev, I. and Audibert, J.-Y. (2011). Data-driven crowd analysis in videos, *2011 International Conference on Computer Vision, Barcelona, Spain*, pp. 1235–1242, DOI: 10.1109/ICCV.2011.6126374.

Shao, J., Change Loy, C. and Wang, X. (2014). Scene-independent group profiling in crowd, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA*, pp. 2219–2226, DOI: 10.1109/CVPR.2014.285.

Tang, S., Andriluka, M., Milan, A., Schindler, K., Roth, S. and Schiele, B. (2013). Learning people detectors for tracking in crowded scenes, *Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia*, pp. 1049–1056, DOI: 10.1109/ICCV.2013.134.

Wang, Q., Chen, M. and Li, X. (2017). Quantifying and detecting collective motion by manifold learning, *AAAI Conference on Artificial Intelligence, San Francisco, USA*, pp. 4292–4298, DOI: 10.5555/3298023.3298190.

Wang, Q., Chen, M., Nie, F. and Li, X. (2020a). Detecting coherent groups in crowd scenes by multiview clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42**(1): 46–58, DOI: 10.1109/TPAMI.2018.2875002.

Wang, Q., Gao, J., Lin, W. and Li, X. (2020b). NWPU-crowd: A large-scale benchmark for crowd counting and localization, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**(6): 2141–2149, DOI: 10.1109/TPAMI.2020.3013269.

Wang, Q., Gao, J., Lin, W. and Yuan, Y. (2020c). Pixel-wise crowd understanding via synthetic data, *International Journal of Computer Vision* **129**(1): 225–245, DOI: 10.1007/s11263-020-01365-4.

Wen, L., Lei, Z., Lyu, S., Li, S. Z. and Yang, M.-H. (2016). Exploiting hierarchical dense structures on hypergraphs for multi-object tracking, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **38**(10): 1983–1996, DOI: 10.1109/TPAMI.2015.2509979.

Yang, B. and Nevatia, R. (2014). Multi-target tracking by online learning a CRF model of appearance and motion patterns, *International Journal of Computer Vision* **107**(2): 203–217, DOI: 10.1007/s11263-013-0666-4.

Yu, H., Zhou, Y., Simmons, J., Przybyla, C.P., Lin, Y., Fan, X., Mi, Y. and Wang, S. (2016). Groupwise tracking of crowded similar-appearance targets from low-continuity image sequences, *IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA*, pp. 952–960, DOI: 10.1109/CVPR.2016.109.

Zhang, L., He, Z., Gu, M. and Yu, H. (2018). Crowd segmentation method based on trajectory tracking and prior knowledge learning, *Arabian Journal for Science and Engineering* **43**(12): 7143–7152, DOI: 10.1007/s13369-017-2995-z.

Zhang, S., Wang, J., Wang, Z., Gong, Y. and Liu, Y. (2015). Multi-target tracking by learning local-to-global trajectory models, *Pattern Recognition* **48**(2): 580–590, DOI: 10.1016/j.patcog.2014.08.013.

Zhou, X., Zhuo, J. and Krahenbuhl, P. (2019). Bottom-up object detection by grouping extreme and center points, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA*, pp. 850–859, DOI: 10.1109/CVPR.2019.00094.

Zhu, F., Wang, X. and Yu, N. (2014). Crowd tracking with dynamic evolution of group structures, *European Conference on Computer Vision, Zurich, Switzerland*, pp. 139–154, DOI: 10.1007/978-3-319-10599-4_10.

Zhu, F., Wang, X. and Yu, N. (2018). Crowd tracking by group structure evolution, *IEEE Transactions on Circuits and Systems for Video Technology* **28**(3): 772–786, DOI: 10.1109/TCSVT.2016.2615460.

Zitouni, M.S., Bhaskar, H. and Al-Mualla, M.E. (2016). Robust background modeling and foreground detection using dynamic textures, *International Conference on Computer Vision Theory and Applications (VISIGRAPP'16), Rome, Italy*, pp. 403–410, DOI: 10.5220/0005724204030410.

Zitouni, M.S., Sluzek, A. and Bhaskar, H. (2019a). CNN-based analysis of crowd structure using automatically annotated training data, *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Taipei, Taiwan*, pp. 1–8, DOI: 10.1109/AVSS.2019.8909846.

Zitouni, M.S., Sluzek, A. and Bhaskar, H. (2019b). Visual analysis of socio-cognitive crowd behaviors for surveillance: A survey and categorization of trends and methods, *Engineering Applications of Artificial Intelligence* **82**: 294–312, DOI: 10.1016/j.engappai.2019.04.012.

Zitouni, M.S., Sluzek, A. and Bhaskar, H. (2020). Towards understanding socio-cognitive behaviors of crowds from visual surveillance data, *Multimedia Tools and Applications* **79**(3): 1781–1799, DOI: 10.1007/s11042-019-08201-z.

M. Sami Zitouni received his PhD and MSc degrees in electrical and computer engineering in 2019 and 2015, respectively, from Khalifa University, Abu Dhabi, UAE. He conducted his studies as part of the KU Center for Autonomous Robotic Systems (KUCARS) and the Visual Signal Analysis and Processing Center (VSAP). He is currently a post-doctoral fellow in biomedical engineering at Khalifa University. His research interests include artificial intelligence, machine learning applications, affective computing, computer vision, signal processing, and embedded systems.

Andrzej Śluzek received his MSc, PhD and DSc degrees from the Warsaw University of Technology. Currently, he is a full professor in the Institute of Information Technology (Department of Artificial Intelligence) of the Warsaw University of Life Sciences (SGGW). From 1992 to 2011 he worked for Nanyang Technological University (School of Computer Science and Engineering) in Singapore, and from 1994 to 2010 he was a deputy director of the Robotic Research Centre of the same university (concurrent appointment). From 2011 to 2020 he was an associate professor at Khalifa University (Department of Electrical Engineering and Computer Science) in Abu Dhabi, UAE. His research interests include machine vision, intelligent robotics and selected aspects of digital signal processing and digital systems.

Received: 16 May 2021
Revised: 14 August 2021
Re-revised: 23 September 2021
Accepted: 19 October 2021